Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples

Alexander Keller[1,2,$], Nadja Danner[1], Gudrun Grimmer[1,2], Markus Ankenbrand[3], Katharina von der Ohe[4], Werner von der Ohe[4], Simone Rost[5], Stephan Härtel[1], Ingolf Steffan-Dewenter[1]

[1] Department of Animal Ecology and Tropical Biology, University of Wuerzburg, Biocenter, Am Hubland, Wuerzburg, Germany

[2] DNA Analytics Core Facility, University of Wuerzburg, Biocenter, Am Hubland, Wuerzburg, Germany

[3] Department of Bioinformatics, University of Wuerzburg, Biocenter, Am Hubland, Wuerzburg, Germany

[4] LAVES Institut für Bienenkunde, Herzogin-Eleonore-Allee 5, Celle, Germany

[5] Department of Human Genetics, University of Wuerzburg, Biocenter, Am Hubland, Wuerzburg, Germany

Keywords: DNA barcoding, high-throughput, internal transcribed spacer 2, ITS2, molecular ecology, meta-barcoding, pollination, plant-pollinator interactions, phylotyping

[$] Corresponding author: Alexander Keller, Department of Animal Ecology and Tropical Biology, Biocenter, Am Hubland, 97074 Würzburg, Germany, eMail: a.keller@biozentrum.uni-wuerzburg.de, Fax: 0049 931 31 84352

Short title: Evaluating NGS-based palynology

Abstract

The identification of pollen plays an important role in ecology, palaeo-climatology, honey quality control and other areas. Currently, expert knowledge and reference collections are essential to identify pollen origin through light microscopy. Pollen identification through molecular sequencing and DNA barcoding has been proposed as an alternative approach, but the assessment of mixed pollen samples originating from multiple plant species is still a tedious and error-prone task. Next-generation sequencing has been proposed to avoid this hindrance. In this study we assessed mixed pollen probes through next-generation sequencing of amplicons from the highly variable species-specific internal transcribed spacer 2 region of the nuclear ribosomal DNA. Further, we developed a bioinformatical workflow to analyse these high-throughput data with a newly created reference database. To evaluate the feasibility, we compared results from classical identification based on light microscopy from the same samples with our sequencing results. We assessed in total 16 mixed pollen samples, 14 originated from honey bee colonies and two from solitary bee nests. The sequencing technique resulted in higher taxon richness (deeper assignments and more identified taxa) compared to light microscopy. Abundance estimations from sequencing data were significantly correlated with counted abundances through light microscopy. Simulation analyses of taxon specificity and sensitivity indicate that 96% of taxa present in the database are correctly identifiable at the genus level and 70% at the species level. Next-generation sequencing thus presents a useful and efficient workflow to identify pollen at the genus and species level without requiring specialized palynological expert knowledge.

Introduction

Palynology, the scientific study of pollen and identification of its origin, plays an important role in studying mechanisms of plant-pollinator interactions (Wilcock and Neiland, 2002), resource use of flower-visiting animals (Kleijn and Raemakers, 2008; Wcislo and Cane, 1996) and climate-related variation of plant communities through time (Marchant et al., 2001; Sugita, 1994; Tzedakis, 1993). Pollen grains often display a species-specific morphology with diverse structure and sculpture. However, it remains difficult to delineate between closely related species when using light microscopy (Mullins and Emberlin, 1997). As a result, many pollen types are simply grouped at genus or family level (Davies and Fall, 2001) and data analyses on pollen diversity are strongly limited (Bagella et al., 2013). DNA barcoding, i.e. to identify and classify organisms according to a nucleotide sequence was often and successfully applied to all major groups of organisms, also plants including pollen (Chen et al., 2010; Hebert et al., 2003; Zhou et al., 2007). Accordingly, molecular tools to analyze pollens have also substantially increased in their application and show great potential especially with difficult, also fossil taxa and those with low taxonomic knowledge (Bennett and Parducci, 2006; Wilson et al., 2010; Zhou et al., 2007).

It is further a promising new approach in ecology to directly determine the diversity of organisms in environmental samples (Sheffield et al., 2009; Valentini et al., 2009), i.e. samples that represent a mixture of species, e.g. faeces, soil or pollen collections, for which identification with classical methods is difficult or incomplete (Wilson et al., 2010). To analyze mixed sets of pollens originating from different plant organisms with DNA barcoding however is still a tedious and error-prone task, requiring manual separation of pollens to taxa, each to be amplified and sequenced individually. Studies evaluating applicability of high-throughput techniques to pollen materials are currently lacking (Taylor and Harris, 2012; Wilson et al., 2010) or are restricted to specific investigations using quantitative real time polymerase chain reaction (qrtPCR) where prior information about present organisms is required (Agodi et al., 2006; Schnell et al., 2010). Palynology would therefore benefit from species-level determination from mixed samples, larger counts, higher processing speed, improved objectivity, and automation to be attractive for large scale studies (Stillman and Flenley, 1996). Molecular methods based on high-throughput DNA-sequencing could provide the requested features to extent and improve classical pollen determination. Valentini et al. (2010) proposed next-generation sequencing (NGS) as a suitable method for this task. We agree with this idea and thus evaluated in this study the performance and reliability of the new sequencing and bioinformatical strategies by directly comparing it with data obtained by light microscopy.

Specifically we address the following challenges that emerge in DNA barcoding with mixed pollen samples. (1) A laboratory routine has to be defined which can be applied to all major plant clades, requiring universality of amplification priming regions and adequate length to be suitable for next-generation sequencing while holding enough sequence variation to differ between species. This routine includes DNA extraction, amplification, sample multiplexing, library preparation, sequencing with high-throughput devices and raw-data cleanup. Also, (2) a mapping algorithm has to be developed which adequately maps obtained sequences in their full length to references, preferably in a hierarchical progression with confidence values for each level of taxonomy. Further, this algorithm has to be with good performance to be able to process high-throughput data on a standard desktop computer and produce results in reasonable time. (3) A comprehensive reference database is required to derive the desired taxonomic annotations.

Several genetic marker regions have been proposed for DNA barcoding in plants that match the requirements, foremost presence and feasibility to be amplified in all investigated taxa, as well as low intra-specific but high inter-specific variability to succeed in being species-specific (Chen et al., 2010; Hebert et al., 2003; Hollingsworth et al., 2011; Zhou et al., 2007). In this study, we use the internal transcribed spacer 2 (ITS2) region, which has been shown to be suitable as a barcode for plants (92.7% successful identifications in 6,600 samples, Chen et al., 2010; Buchheim et al., 2011). Also the enclosed genetic regions (5.8S and 28S) are highly conserved throughout the eukaryotes. Thus an universal primer for the analysis of probes consisting of multiple organisms is applicable with a low risk to exclude taxa from amplification (Chen et al., 2010; Keller et al., 2009; White et al., 1990). A further reason for choosing this marker is that a comprehensive ITS2 database already exists (Koetschan et al., 2010) helping to prepare reference sequences suitable for our needs.

We approached the targeted tasks by combining and adapting existing molecular and bioinformatical tools to develop new functionalities for DNA barcoding of pollen samples that consist of multiple taxa. We then evaluated the performance and quality of the molecular and bioinformatical workflow by comparing our results with data from classical light microscopy identification of pollen samples. Further, we tested the applicability for samples with low pollen contents and performed computer-based simulations to validate that the bioinformatical classification pipeline is trustable.

Materials and Methods

Pollen collection

The honey-bee pollen samples were collected in twelve different landscapes in the region around Bayreuth, Germany. The distance between landscapes was at least 3 km leading to diversified pollen inputs depending on the surrounding floral resources. In the centre of each landscape we established a honey bee colony (*Apis mellifera carnica* L.) with a pollen trap in front of the hive entrance. Returning foragers had to pass a 5 mm grid taking off the pollen loads from their hind legs. From 21.07.2009-12.08.2009 every one to three days accumulated pollen loads were removed from the traps and stored as individual samples at -18 °C until the end of the sampling period. Pollen samples were dried at 30 °C for one week. Further, to assess variability in resource use of honeybees at one location, samples from three colonies located at the same study site were separately analysed (in the following designated as Samples 12a, 12b and 12c). From each of the fourteen samples (one per colony) 20% of the collected pollen were randomly taken and mixed for further analyses.

We performed NGS as well as microscopic assessment of the samples. The samples were split into independent aliquots for these separate, blinded analyses. NGS was performed by AK, GG and MA, whereas the samples were classified through classical light microscopy by ND with expert guidance by KvO, without knowledge of the other group's results.

Two further pollen samples were obtained from solitary bee nests (*Osmia bicornis* L.) in October 2012 by swabbing the cell walls with cotton buds (Keller et al., 2013). In contrast to the relatively pure pollen samples obtained from honey bees, this experiment reflects samples strongly contaminated with nest building materials (soil) and faeces, challenging to analyse with traditional methods. Solitary bee samples were thus only processed with NGS.

Classical pollen identification

Pollen samples were first analyzed using light microscopy in the LAVES Institut für Bienenkunde in Celle, Germany. For the microscopic pollen determination, 10mg pollen loads of each sample were homogenized in 50ml demineralized water with a magnetic stirrer for one hour. 15 µl of the solution and 30 µl demineralized water were transferred to a slide, distributed equally over an area of the size of a cover glass and embedded in glycerin gelatin after complete dehydration following the method of Behm et al. (1996). From each sample 500 randomly selected pollen grains were determined on genus level and where possible to species level. Very rarely occurring pollen types were not determined (Behm et al., 1996).

Molecular pollen identification

Second pollen identification was done by DNA barcoding of the ITS2 region. The main working steps described below were DNA extraction, amplification, sequencing, bioinformatic cleanup and taxonomic classification.

DNA extraction, amplification and sequencing: For each sample, 2 g of pollens were added to 4 ml of bidest H_2O and homogenized with an electronic pistil within a plastic tube. Of this emulsion, 200 µl (equaling approximately 50 mg of pollens) were taken for the following extraction. We grinded the aliquot with the TissueLyser LT (Qiagen, Hilden, Germany) and extracted DNA using the Machery-Nagel (Düren, Germany) NucleoSpin Food Kit. We followed the special supplementary guidelines for pollen samples provided by the manufacturer. For polymerase chain reaction (PCR) amplification we used the primers S2F and ITS4R originally designed by Chen et al. (2010) and White et al. (1990) to span a mean region of approximately 350bp. This covers the complete ITS2 region. We adapted those primers to match 454 sequencing purposes and multiplexing by adding the 454 specific Adapters A and B, the linker key, and a variable multiplex identifier (MID). Thus the forward "fusion" primer was 5'-CGT ATC GCC TCC CTC GCG CCA TCA GAT GCG ATA CTT GGT GTG AAT -3' and the reverse "fusion" primer 5'-5´CTA TGC GCC TTG CCA GCC CGC TCA GXX XXX XXX XXT CCT CCG CTT ATT GAT ATG C-3', where the X-region designates a variable MID. In total, 16 MIDs were taken from the official Roche technical bulletin (454 Sequencing Technical Bulletin No. 005-2009, April 2009) to be able to process all our samples with one sequencing chip.

PCR reaction mixes consisted of 0.25 µl of each forward and reverse primer (each 30µM molar), 3 µl of template DNA and 25µl of Phusion High-Fidelity DNA polymerase PCR 2x MasterMix (Thermo Scientific, Waltham, MA, USA). Bidest H_2O was added to a reaction volume of 50 µl. Samples were initially denaturated at 94 °C for 4 min, then amplified by using 37 cycles of 95 °C for 40 s, 49 °C for 40 s and 72 °C for 40 s. A final extension (72 °C) of 5 min was added at the end of the program to ensure complete amplification. All samples were amplified in ten separate aliquots to reduce random effects on the community during PCR amplification (Fierer et al., 2008). PCR amplicons of these ten replicates were combined, gel-electrophoresed, trimmed for amplicon length and cleaned with the HiYield PCR Clean-up Kit (Real Biotech Corporation, Banqiao City, Taiwan) according to the manufacturers description. Cleaned samples were quantified using a Qubit II Flurometer (Invitrogen/Life Technologies, Carlsbad, CA, USA) and the dsDNA High-Sensitivity Assay Kit (also Invitrogen/Life Technologies) as described in the vendors protocol. We used the BioAnalyzer 2200 (Agilent, Santa Clara, CA, USA) with High Sensitivity DNA Chips (also Agilent) for verification of fragment length distributions. Pyrosequencing and library preparation was performed according to the guidelines for the GS junior (Roche, Basel, Switzerland). Sequencing was performed in-house with a GS junior device located in the Department of Human Genetics (University of Würzburg, Germany) with original Roche GS junior titanium chemistry.

Bioinformatic cleanup: Data was demultiplexed into the different samples using the MID adapter sequences and the QIIME software (Caporaso et al., 2010; Kuczynski et al., 2011). During this step, only sequences spanning both priming regions were further used, i.e. only completely sequenced amplicons. Primers, adapters and MIDs were trimmed. Chimeric checking and quality filtering was also performed during this step. We restricted data to high quality reads with a phred score ≥ 27, (Kunin et al., 2010) and no reads with ambiguous characters were included in the following downstream analyses.

Hierarchic classification: Taxonomic assignments were performed with the RDP (Ribosomal Database Project) classifier (Wang et al., 2007) and an ITS2 specific, novel reference set created and evaluated as described below. Further, we applied a bootstrap cut-off at 85% as classification threshold with respect to the maximum f-measure in the training database evaluation (also see below).

Method comparison statistics

Most of the analyses were performed on a generic level, as both methods yielded some taxa only assignable to this level. With a generic analysis all identified taxa were directly comparable. With this data we compared taxon richness and identified species overlaps and differences obtained from the two methods. Rarefaction curves for each plot were generated with R (R Development Core Team, 2010) in the NGS data to evaluate species richness in relation to sequencing depth. Abundance was assessed relatively in percent of total number of reads and in percent of 500 pollen grains (Behm et al., 1996) for NGS and light microscopy, respectively. We used overall and per-plot abundance of these relative accounts to compare between the two methodologies by Pearson’s product moment correlation using R (R Development Core Team, 2010).

Molecular reference database training

Taxonomic classifications with DNA barcodes are currently mostly done via phylogenetic analyses (Buchheim et al., 2011), pairwise alignments with specific reference sequences (Chen et al., 2010) or BLAST searches (Basic Local Alignment Search Tool) (Altschul et al., 1990) in GenBank (Benson et al., 2010) or other nucleotide databases. The first both methods require that prior knowledge about taxonomy is present to select suitable taxa included into the recalculated phylogenetic tree or alignment. This is not feasible for mixed pollen collections, where the included taxa are unknown prior to assessment or stem from very different taxonomic groups. BLAST searches have to be performed very carefully, as hits may include local alignments and identity calculations may thus be based only on parts of the query and reference sequences. Further, the raw output of a BLAST search is often obscured as a lot of hits are not taxonomically annotated or flagged as “environmental samples”. A novel approach to tackle these drawbacks has been proposed with a Bayesian classification algorithm (Wang et al., 2007). It provides hierarchical taxonomic assignments of DNA sequences and is well accepted in the scientific community as especially high throughput analyses profit from the efficiency and accuracy of the algorithm (Caporaso et al., 2010). Currently, the only publicly available training sets are limited to bacterial 16S (Wang et al., 2007) and fungal large ribosomal subunit (Liu et al., 2012).

In this study, a new ITS2 training set was designed for plants. We used the ITS2-Database as an original database which is restricted to structure-validated sequences (Koetschan et al., 2010). All ITS2 sequences matching the taxonomic group "Viridiplantae" and with a sequence length between 200 bp and 400 bp were downloaded, resulting in 73,853 sequences (accessed 3rd March 2013). The taxonomy for each sequence was assigned using the GI (GenBank Identifier) and the corresponding NCBI taxonomy (Federhen, 2012) by Perl scripting and reformated to be usable with the python script "assign taxonomy.py" of the QIIME (Caporaso et al., 2010) package. Additionally, RDP required formats of these preprocessed files were generated. Training was performed with the RDP classifier v2.2 (Wang et al., 2007) as implemented in QIIME. Before training of the final set, we evaluated the performance by varying several parameters of the underlying data to maximize effectiveness and allow quality estimations of the assignments as described in the following.

Pre-clustering evaluation: Due to intraspecific variation (Song et al., 2012) and sequencing errors in the underlying data (Kunin et al., 2010), pre-clustering of reference sequences prior to training may prove useful to increase reliability of the results (Lan et al., 2012). Thus, from the full data-set we generated eleven separate training sets differing in the pre-clustering threshold of sequences before the actual training. Clusters of sequences were generated at identity levels of 90%, 91% … 100%, and only the most abundant sequence of each cluster was picked. This also generated an even distribution of taxonomic units in the sets. To assess the assignment quality and depth, each sequence was reclassified to the training set. Then starting from the root of the taxonomy of each sequence, every taxonomic level of the assignment was compared to the correct taxonomy. If the bootstrap of an assignment was less than 0.8, the level (and all sub-levels) was considered as unassignable. If there was a mismatch between assigned taxonomy and expected taxonomy, the number of remaining sub-levels (plus one), was called erroneous levels. The number of assigned levels before the first mismatch or unassignable level was called correct levels.

Cut-off and assignment quality evaluation: To estimate assignment qualities, the test and training data had to be distinct sets. Further, we wanted to evaluate the effectiveness to identify “new species” that do not have representatives in the training data (Lan et al., 2012). The complete ITS2 reference data set was thus for testing purposes artificially split into three sets representing “training data”, “test data A” with references, and “test data B” without references. This was achieved by the following procedure: species with multiple sequences were separated into “test data A” (one sequence) as well as “training data“ (remaining sequences). Species with only a single deposited sequence were assigned to category “test data B”. For this evaluation purpose, the algorithm was trained only with the set “training data“ (36,418 sequences). According to the measures for the RDP classifier evaluation performed by Lan et al. (2012) for the original 16S dataset we estimated the number of “true positive” (TP) and “false negative” (FN) assignments by classifying sequences of “test data A” (10,635 sequences), where references were present in the “training data”, Only correct assignments were considered as TP, whereas wrong assignments (to a different species) were added to the list of FNs. Similarly, we classified sequences of “test data B” (26,800 sequences) to determine the number of “true negative” (TN) and “false positive”(FP) hits. With that, we calculated sensitivity $SN = \frac{TP}{TP+FN}$ to identify existing taxa and specificity $SP = \frac{TN}{TN+FP}$ to leave sequences without references unclassified. Using these split data-sets, we were able to estimate SN at species and genus level, whereas SP was only assessable at the species level. We optimized our assignment bootstrap value for classification by maximizing the f-measure as the harmonic mean of sensitivity and specificity at species level $= \frac{2*SN*SP}{SN+SP}$.

Results

Pollen high-throughput sequencing and classification

In total, our study produced 14,924 raw sequences for pollen samples passing Roche's quality filtering of the 454 junior sequencing device. Of these, 9,310 ITS2 sequences matched our extended quality standards. The remainders were dismissed as too short (<200 bp), with low quality score (<27), excessive homopolymers (>5 bp), chimeric or mismatches in primer regions (Caporaso et al., 2010; Kunin et al., 2010). After removal of adapters and primers, mean sequence length was 348,3 bp (± 28 bp standard deviation), spanning the complete ITS2 region. Individual samples comprised 219-1,179 reads, with mean read length of 330,5 bp – 363,9 bp (± 3,8 bp – 68,2 bp standard deviation). Beside plant sequences, we also found several fungal sequences, belonging to *Issatchenkia occidentalis*, *Cochliobolus sativus*, *Phoma* sp. and *Lewia infectoria*, which are regularly inhabiting or infecting plant tissues.

Honey bee pollen samples

For the samples collected by honey-bees, 98.9% of all reads were assignable to genus level with a bootstrap confidence higher or equal than 0.85. At the species level we were able to classify 61.6% of our reads using the same bootstrap cut-off. Reducing the filter's required sequence length to 150 bp did not produce any new classifiable plant taxa. Taxon richness was not correlated with the number of reads within a sample (Pearson's correlation, r = -0.099, df = 12, t = -0.3453, p-value > 0.05). Rarefaction showed that we reached a plateau regarding genera richness in all samples (Fig. 1A). These observations suggest that the sequencing depth was adequate to assess the underlying taxon richness.

We identified a total of 29 different genera of 16 families when we combined the results from molecular sequencing and microscopy (Tab. 1). Further, 24 taxa were also identifiable at the species level. With NGS we found 13 genera that were not identified through microscopy, whereas four genera (*Heracleum, Carduus, Phacelia, Convolvulus*) that were identified by light microscopy were missing in the NGS results although having references in the database. One genus (*Vitis*) had no trustable reference sequence in the database and was thus also not identifiable with the NGS method.

From the phenology of the pollens and presence at plots, we assume that a misidentification of very similar pollens happened with light microscopy which was revealed by NGS: *Tanacetum* and *Scorzoneroides* were both manually misclassified as *Taraxacum*. We observed higher intra-generic taxon richness for *Trifolium*, *Hypochaeris*, *Chamerion* through NGS, yet lesser in *Centaurea* (Fig. 1B). Improvement of the taxonomic assignment was found in four genera, where species levels were obtainable only through NGS. However, *Helianthus* was only classified at genus level, whereas microscopy was able to identify it as *Helianthus annuus*.

Based on NGS data, taxon richness within the samples ranged from 4 to 12 taxa that were at least classifiable at genus level (Fig. 1B). Correspondingly, diversity ranged from 4 to 12 taxa for the microscopy assessment. Pollen diversity collected by the three colonies from site twelve was 12, 10 and 12 taxa, respectively. The compositional profile was similar for the dominant pollen taxa in all three samples, but still showed considerable variation (Fig. 1B).

Over all samples, we found a strong correlation of abundance estimations between the two identification methods (Pearson's correlation, $r = 0.86$, $t = 8.71$, $df = 26$, $p < 0.001$***, Fig. 2). This relationship is also reflected on a per plot basis, yet with lower correlation coefficient (Pearson's correlation, $r = 0.66$, $t = 17.36$, $df = 390$, $p < 0.001$***). These results indicate that the abundance estimates of taxa within plots show relatively high similarity between the two methods.

Pollens in solitary bee nests

Pollen samples from both solitary bee nests were successfully processed with 100% of reads identifiable at genus level despite high contamination of the samples with nesting materials and faeces. Both samples harbored *Brassica* sp. and *Dioscorea* sp.

pollens, the latter most likely *Dioscorea* (*Tamus*) *communis* as the only representative of the Dioscoreaceae present in the sampling region.

Molecular reference database training

Pre-clustering of data prior to training of the RDP classifier did not improve the overall performance of classifications (Fig. 3). This was the case both for depth of the assignment as well as the mean number of incorrectly assigned levels, which respectively increase and decrease with higher pre-clustering thresholds. We thus used a cut-off at 100% sequence identity, which equals unique sequences, for the final training set. With that, of the 73,853 tested database sequences, 55,028 were positively identifiable at the species and further 10,518 at the genus level. Surprisingly, 6,104 sequences were assignable only to phylum level. They likely represent contaminations in the reference database.

Regarding determination of the optimal cut-off threshold, specificity and sensitivity of the novel/known classifications are shown with their dependency of the bootstrap in Fig. 4. The best classification by means of f-measure is achieved with a bootstrap cutoff of 0.85. Specificity and sensitivity at this threshold for species level were both approximately 70%. At the genus level, sensitivity to correctly identify a genus increased to 96%. We thus recommend this threshold when using the RDP classifier with the generated training data.

Currently, all sequences in the reference data-set accumulate to 37,435 different plant species and 6,162 genera according to NCBI taxonomy (Federhen, 2012). The complete reference dataset is available for download and public usage at http://www.dna-analytics.biozentrum.uni-wuerzburg.de.

Discussion

The demand for methods to identify pollen samples at a high-throughput level is increasing for many applications in ecology and paleo-climatology (Bennett and Parducci, 2006; Sheffield et al., 2009; Taylor and Harris, 2012; Valentini et al., 2009;

Wilson et al., 2010; Zhou et al., 2007). DNA barcoding is a frequently and successfully applied method, yet pollens of mixed samples originating from more than one source are currently not assessable through standard methods. Valentini et al. (2010) proposed that next-generation sequencing may counter this deficiency, i.e. to investigate such mixed samples by identifying all included plant organisms together without manual separation. The goals of this study were thus to develop, and moreover evaluate, a molecular laboratory procedure and bioinformatical analysis for such a task. The complete workflow was applied to pollen samples from two different studies (in total 16 samples). The resulting gene sequences allowed to successfully identify taxon richness and abundance of the underlying samples. The resulting taxonomic resolution is similar or better than results from classical light microscopy. Details of the performance of each individual step of the workflow and the resulting methodological and biological relevance are discussed in the following.

High-throughput pollen sequencing

In general, our laboratory workflow was suitable in processing mixed pollen probes through next-generation sequencing. However, quality filtering according to our rigorous restrictions reduced the obtained sequences from approximately 15,000 sequences to 10,000. Most of them were removed due to failure to include both primer regions and/or multiplex identifier due to low quality scores towards the end of sequences or short read lengths (Caporaso et al., 2010). The first indicates that a large proportion of reads was not fully sequenced with sufficient quality, whereas the latter shows that the primers also amplified shorter fragments than the intended plant ITS2 region. Not fully sequenced reads are a technical issue that is regularly improved by increase of read length and quality through new generations of sequencing devices and chemistry (Metzker, 2009). Improvements are also expectable by applying paired-end strategies, as quality near the ends will increase, or to use technologies with general lower sequencing error rates. Shorter, fully sequenced sequences are project specific problems, but also expectable: as a drawback of universal primers, they will as well amplify fungal ITS2 (White et al., 1990) ranging from ~100 to 250 bp and even other eukaryotic protists with far shorter ITS2 regions (Keller et al., 2009). Further, the existence of non-functional pseudo-genes is known (Harpke and Peterson, 2008). Thus studies investigating plant ITS2 sequences should account for a sufficient overhead of estimated sequences per sample during project design due to sequencing technology and potential contamination through unwanted organisms (Parameswaran et al., 2007). The remaining high quality reads showed a high proportion of classifiable sequences (~99%), whereas reduction of the minimum sequence length had no impact on plant species diversity. Both observations suggest that the filters are adequate to concentrate on the data of interest, i.e. plant sequences.

Classification pipeline

To be able to use the RDP classifier (Wang et al., 2007) for taxonomic assignments with plants and with the ITS2 marker, we re-trained the algorithm with structurally verified sequences obtained from the ITS2 database (Koetschan et al., 2010). The underlying dataset incorporates more than 70,000 different plant sequences and represents a cross-section throughout the Viridiplantae. Sequences originate from all biogeographic regions of the world since the primary database is GenBank (Benson et al., 2010). Currently, all sequences in the reference data-set accumulate to 37,435 different plant species and 6,162 genera according to NCBI taxonomy (Federhen, 2012). Exemplarily for the data analysed in this study, the dataset covers 79% of all vascular plant genera and 54% of species known to exist within the Federal state Bavaria, Germany, where our samples were obtained (comprehensive plant database http://www.bayernflora.de, accessed 6th November 2013, Staatliche Naturwissenschaftliche Sammlungen Bayern, 2013). As 99% of reads were classifiable to genus level and only one genus (*Vitis*) of the assessed 29 genera in total was missing in the reference database, most of abundant and bee relevant plant genera seem to be included. Further, the classifier's dataset is updateable to match the constantly increasing numbers of sequences deposited in GenBank and the ITS2 database in the future (Wang et al., 2007).

In the computational evaluation of database and classifier for an ITS2 dataset, we obtained values comparable to those of existing datasets published for bacteria (Wang et al., 2007) and fungi (Liu et al., 2012). Taxonomic classifications performed best regarding sensitivity, i.e. to identify taxa existing in the database, and specificity, i.e. to restrain from classifying organisms without references, at a bootstrap threshold level of approximately 0.85 (Lan et al., 2012). Species and genus level sensitivity to correctly identify sequences with this bootstrap were 70% and 96%, respectively. This is similar to the classifier's preferred level used to classify microbial organisms (0.80, Lan et al., 2012; Wang et al., 2007). From a technical perspective it is thus valid to apply the classification algorithm also for ITS2 sequences of plants.

Comparison of assessment methods

Using next-generation sequencing, we were clearly able to improve palynology diversity assessments in comparison with traditional optical microscopy. This appears in novel taxa that were identified, as well as improvement of classification of taxa and better possibilities to distinguish species within a genus. Further, some misidentifications of pollen through microscopy were revealed that were caused by very similar morphological appearance of closely related species. Also, molecular assessments were successful for solitary bee nest samples, where swabs included pollens as well as contaminating materials. Sequencing assessments were repeatable, identifying similar diversity in samples obtained from different bee colonies placed within the same landscape.

However, using the high-throughput approach we also encountered limitations, which are partly related to the data used for training of the classifier. Regarding the Vitaceae, the ITS2 database is currently lacking trustable reference sequences. We validated the only existing sequence, which was considerably short (~200 bp) and derived from a whole genome shotgun sequencing study (assembled sequence from short length reads, GenBank ID: AM462492.2, Velasco et al., 2007). Due to intra-genomic variation of the ITS2 (Song et al., 2012), we assume the assembly yielded a consensus, stacked ITS2 sequence, unusable for barcoding purposes or that a non-ITS2 region was falsely identified as such by the ITS2 database annotation algorithm (Keller et al., 2009). We therefore dismissed the sequence as missing within the reference database. In general, taxa missing or with inadequate sequences in the underlying database are not identifiable. As shown exemplarily for the geographic region Bavaria, 22% of known plant genera are missing and thus the current coverage is far from complete (Staatliche Naturwissenschaftliche Sammlungen Bayern, 2013). Also, valid sequences with wrong taxonomic annotations may lead to mis-training of the classification model regarding the respective taxa (Bridge et al., 2003). This is exemplified by a proportion of sequences re-classified in the evaluation to a different phylum, suggesting wrong taxonomic annotation of GenBank database sequences. To address limitations of the underlying database (missing or misclassified sequences) in a given research question, we suggest that applied studies should consider also reviewing one cross-section pool of all samples in parallel through optical means to verify the overall richness of taxa relevant for the study. This will also maintain comparability between studies applying traditional and molecular approaches. Despite these database-specific drawbacks, the classifier produced taxonomic assignments that are congruent with light microscopy, and thus corroborating the positive technical evaluation of the pipeline above with a direct comparison of biological data.

Abundance estimations of both methods showed a strong correlation, suggesting that abundance estimates based on high-throughput sequencing regarding high or low sequence frequency of taxa within the sample are valid. In our study, we took care to reduce amplification biases through PCR with ten aliquots of each sample simultaneously (typical in microbiota studies: three, Fierer et al., 2008) and a low number of amplification cycles (Suzuki and Giovannoni, 1996). Still, abundances retained from PCR amplified DNA samples have to be regarded critically, as amplification biases through priming preference of specific taxonomic groups, random effects and the exponential nature of the amplification process are not excludable (Spooner, 2009; Suzuki and Giovannoni, 1996). Abundances are thus likely better interpreted categorical (e.g. high abundance, low abundance) than with linear association. With the advent of increased sequencing throughput and third-generation single molecule sequencers without need for amplification (Metzker, 2009; Roberts et al., 2013), improved abundance estimations by sequencing are likely in the near future.

Expenses per sample were almost equal for both applied methods when considering time consumption and consumables. As the trend of sequencing technologies goes rapidly toward higher throughput and resulting multiplexing possibilities (Kozich et al., 2013; Metzker, 2009), we expect price efficiency per sample with next-generation sequencing to outpace optical assessments in the near future.

Fields of application

Various applications arise for the proposed method. These include studies of pollen material from various origins, including plants themselves, pollinators, soil samples and wind collections. The results of such assessments are of great importance in identifying the diversity and specialization of plant-pollinator interaction networks (Bosch et al., 2009) and also in supporting agricultural and ecological management decisions (e.g. Girard et al., 2012; Odoux et al., 2012). Further, paleo-ecological and climate-change associated studies investigating fossil pollens may also largely profit (Bennett and Parducci, 2006).

Special attention is currently required in quality control of honey-bee products, including the geographical origin, correct labeling of different varieties based on the used floral resources and detection of contaminations from genetically modified (GM) crops (Hemmer, 1997; Picard-Nizou et al., 1995). As pollen is naturally incorporated into honey and protocols to isolate them are common usage (Sowunmi, 1976), high-throughput sequencing and classification may contribute largely to this endeavor by facilitating the analytical process and inclusion of references from plant taxa throughout the world (Ruoff et al., 2007; Sowunmi, 1976).

Furthermore, the methodology may be equivalently applied to other questions not only related to pollens. Other target samples are naturally occurring communities of plants, (e.g. green algae), or artificially mixed probes of plant tissue fragments (Schlumbaum et al., 2008). As the primers used in this study also efficiently amplify fungal ITS2 sequences, ancillary information is automatically gained about this group including pathogens as *Ascosphaera* spp. that may be present in collected pollen samples and vectorised through harvesting flights of worker bees (Gilliam, 1990; White et al., 1990).

Conclusions

Expert knowledge is essential to identify pollens adequately through traditional light microscopy and taxonomic expertise is also often restricted to specific plant groups or geographical regions. Further, mixed samples of pollens from several plant origins present a problem in current palynology. With this study we evaluated next-generation sequencing to approach pollen assessments through molecular techniques including their bioinformatical analysis. The analytical pipeline is designed for high-throughput data, but also adaptable to single sequences. It is a useful technique broadening the assessment capabilities from expert labs to all workgroups with access to standard molecular laboratory equipment. Further, our results show that this assessment method improves the standard technique with regard to taxonomical deepness, overall diversity and rectifying misidentifications.

Acknowledgements

We greatly appreciate the help of staff from the Department of Human Genetics (University of Würzburg, Germany), especially C. Müller-Reible and T. Haaf for placing the 454 device at our disposal for this study. We also thank F. Förster (Department of Bioinformatics, University of Würzburg, Germany), for providing a useful taxonomy Perl module for generation of the training sets and two anonymous reviewers for their valuable comments.

Data Accessibility

Sequences have been deposited at the ENA:SRA (https://www.ebi.ac.uk/ena) and are accessible under study accession number PRJEB5016. The used training set alongside installation and application notes is available for download at http://www.dna-analytics.biozentrum.uni-wuerzburg.de.

References

Agodi, A., Barchitta, M., Grillo, A., Sciacca, S. (2006) Detection of genetically modified DNA sequences in milk from the Italian market. International Journal of Hygiene and Environmental Health, **209**(1), 81-88.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. Journal of Molecular Biology, **215**(3), 403-410.

Bagella, S., Satta, A., Floris, I., Caria, M.C., Rossetti, I., Podani, J. (2013) Effects of plant community composition and flowering phenology on honeybee foraging in Mediterranean sylvo-pastoral systems. Applied Vegetation Science.

Behm, F., von der Ohe, K., Henrich, W. (1996) Zuverlässigkeit der Pollenanalyse von Honig: Bestimmung der Pollenhäufigkeit. Deutsche Lebensmittel-Rundschau, **92**(6), 183-188.

Bennett, K.D. and Parducci, L. (2006) DNA from pollen: principles and potential. The Holocene, **16**(8), 1031-1034.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W. (2010) GenBank. Nucleic Acids Research, **38**(suppl 1), D46-D51.

Bosch, J., Martín González, A.M., Rodrigo, A., Navarro, D. (2009) Plant–pollinator networks: adding the pollinator's perspective. Ecology Letters, **12**(5), 409-419.

Bridge, P.D., Roberts, P.J., Spooner, B.M., Panchal, G. (2003) On the unreliability of published DNA sequences. New Phytologist, **160**(1), 43-48.

Buchheim, M., Keller, A., Koetschan, C., Forster, F., Merget, B., Wolf, M. (2011) Internal Transcribed Spacer 2 (nu ITS2 rRNA) sequence-structure phylogenetics: towards an automated reconstruction of the green algal tree of life. PloS one, **6**(2).

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I. (2010) QIIME allows analysis of high-throughput community sequencing data. Nature Methods, **7**(5), 335-336.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PloS one, **5**(1), e8613.

Davies, C.P. and Fall, P.L. (2001) Modern pollen precipitation from an elevational transect in central Jordan and its relationship to vegetation. Journal of Biogeography, **28**(10), 1195-1210.

Federhen, S. (2012) The NCBI taxonomy database. Nucleic Acids Research, **40**(D1), D136-D143.

Fierer, N., Hamady, M., Lauber, C.L., Knight, R. (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proceedings of the National Academy of Sciences, **105**(46), 17994-17999.

Gilliam, M. (1990) Chalkbrood disease of honey bees, Apis mellifera, caused by the fungus, *Ascosphaera apis*: A review of past and current research. Vth International Colloquium on Invertebrate Pathology and Microbial Control, Adelaide, Australia, 398-402.

Girard, M., Chagnon, M., Fournier, V. (2012) Pollen diversity collected by honey bees in the vicinity of Vaccinium spp. crops and its importance for colony development 1 1 This article is part of a Special Issue entitled "Pollination biology research in Canada: Perspectives on a mutualism at different scales". Botany, **90**(7), 545-555.

Harpke, D. and Peterson, A. (2008) 5.8 S motifs for the identification of pseudogenic ITS regions. Botany, **86**(3), 300-305.

Hebert, P.D., Cywinska, A., Ball, S.L. (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, **270**(1512), 313-321.

Hemmer, W. (1997) *Foods derived from genetically modified organisms and detection methods* Agency for Biosafety Research and Assessment of Technology Impacts of the Swiss Priority Programme Biotechnology of the Swiss National Science Foundation.

Hollingsworth, P.M., Graham, S.W., Little, D.P. (2011) Choosing and using a plant DNA barcode. PloS one, **6**(5), e19254.

Keller, A., Grimmer, G., Steffan-Dewenter, I. (2013) Diverse microbiota identified in whole intact nest chambers of the red mason bee *Osmia bicornis* (Linnaeus 1758). PloS one, e78296.

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T., Wolf, M. (2009) 5.8 S-28S rRNA interaction and HMM-based ITS2 annotation. Gene, **430**(1-2), 50-57.

Kleijn, D. and Raemakers, I. (2008) A retrospective analysis of pollen host plant use by stable and declining bumble bee species. Ecology, **89**(7), 1811-1823.

Koetschan, C., Forster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R., Muller, T., Wolf, M., Schultz, J. (2010) The ITS2 Database III -sequences and structures for phylogeny. Nucleic Acids Research, **38**(Database issue), D275-279.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology.

Kuczynski, J., Stombaugh, J., Walters, W.A., González, A., Caporaso, J.G., Knight, R. (2011) Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Current Protocols in Bioinformatics, **10**, 1-10.17.

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental Microbiology, **12**(1), 118-123.

Lan, Y., Wang, Q., Cole, J.R., Rosen, G.L. (2012) Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. PloS one, **7**(3), e32491.

Liu, K.-L., Porras-Alfaro, A., Kuske, C.R., Eichorst, S.A., Xie, G. (2012) Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. Applied and Environmental Microbiology, **78**(5), 1523-1533.

Marchant, R., Behling, H., Berrio, J.C., Cleef, A., Duivenvoorden, J., Hooghiemstra, H., Kuhry, P., Melief, B., Geel, B.V., Hammen, T.V.d. (2001) Mid-to Late-Holocene pollen-based biome reconstructions for Colombia. Quaternary Science Reviews, **20**(12), 1289-1308.

Metzker, M.L. (2009) Sequencing technologies—the next generation. Nature Reviews Genetics, **11**(1), 31-46.

Mullins, J. and Emberlin, J. (1997) Sampling pollens. Journal of Aerosol Science, **28**(3), 365-370.

Odoux, J.-F., Feuillet, D., Aupinel, P., Loublier, Y., Tasei, J.-N., Mateescu, C. (2012) Territorial biodiversity and consequences on physico-chemical characteristics of pollen collected by honey bee colonies. Apidologie, **43**(5), 561-575.

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M., Fire, A.Z. (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Research, **35**(19), e130.

Picard-Nizou, A., Pham-Delegue, M., Kerguelen, V., Douault, P., Marilleau, R., Olsen, L., Grison, R., Toppan, A., Masson, C. (1995) Foraging behaviour of honey bees (*Apis mellifera* L.) on transgenic oilseed rape (B*rassica napus* L. var. *oleifera*). Transgenic Research, **4**(4), 270-276.

R Development Core Team. (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria(01/19).

Roberts, R.J., Carneiro, M.O., Schatz, M.C. (2013) The advantages of SMRT sequencing. Genome Biology, **14**(6), 405.

Ruoff, K., Luginbühl, W., Kilchenmann, V., Bosset, J.O., von der Ohe, K., von der Ohe, W., Amadò, R. (2007) Authentication of the botanical origin of honey using profiles of classical measurands and discriminant analysis. Apidologie, **38**(5), 438-452.

Schlumbaum, A., Tensen, M., Jaenicke-Després, V. (2008) Ancient plant DNA in archaeobotany. Vegetation History and Archaeobotany, **17**(2), 233-244.

Schnell, I.B., Fraser, M., Willerslev, E., Gilbert, M.T.P. (2010) Characterisation of insect and plant origins using DNA extracted from small volumes of bee honey. Arthropod-Plant Interactions, **4**(2), 107-116.

Sheffield, C.S., Hebert, P.D., Kevan, P.G., Packer, L. (2009) DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. Molecular Ecology Resources, **9 Suppl s1**, 196-207.

Song, J., Shi, L., Li, D., Sun, Y., Niu, Y., Chen, Z., Luo, H., Pang, X., Sun, Z., Liu, C. (2012) Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. PloS one, **7**(8), e43971.

Sowunmi, M. (1976) The potential value of honey in palaeopalynology and archaeology. Review of Palaeobotany and Palynology, **21**(2), 171-185.

Spooner, D.M. (2009) DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. American Journal of Botany, **96**(6), 1177-1189.

Staatliche Naturwissenschaftliche Sammlungen Bayern. (2013) Botanischer Informationsknoten Bayern.

Stillman, E. and Flenley, J.R. (1996) The needs and prospects for automation in palynology. Quaternary Science Reviews, **15**(1), 1-5.

Sugita, S. (1994) Pollen representation of vegetation in Quaternary sediments: theory and method in patchy vegetation. Journal of Ecology, 881-897.

Suzuki, M.T. and Giovannoni, S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Applied and Environmental Microbiology, **62**(2), 625-630.

Taylor, H.R. and Harris, W.E. (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources, **12**(3), 377-388.

Tzedakis, P. (1993) Long-term tree populations in northwest Greece through multiple Quaternary climatic cycles. Nature, **364**(6436), 437-440.

Valentini, A., Miquel, C., Taberlet, P. (2010) DNA barcoding for honey biodiversity. Diversity, **2**(4), 610-617.

Valentini, A., Pompanon, F., Taberlet, P. (2009) DNA barcoding for ecologists. Trends in Ecology & Evolution, **24**(2), 110-117.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L.M., Vezzulli, S., Reid, J. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PloS one, **2**(12), e1326.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, **73**(16), 5261-5267.

Wcislo, W.T. and Cane, J.H. (1996) Floral resource utilization by solitary bees (Hymenoptera: Apoidea) and exploitation of their stored foods by natural enemies. Annual Review in Entomology, **41**, 257-286.

White, T., Bruns, T., Lee, S., Taylor, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR-protocols: a Guide to Methods and Applications, 315 - 322.

Wilcock, C. and Neiland, R. (2002) Pollination failure in plants: why it happens and when it matters. Trends in Plant Science, **7**(6), 270-277.

Wilson, E.E., Sidhu, C.S., LeVan, K.E., Holway, D.A. (2010) Pollen foraging behaviour of solitary Hawaiian bees revealed through molecular pollen analysis. Molecular Ecology, **19**(21), 4823-4829.

Zhou, L.J., Pei, K.Q., Zhou, B., Ma, K.P. (2007) A molecular approach to species identification of Chenopodiaceae pollen grains in surface soil. American Journal of Botany, **94**(3), 477-481.

Tables

Tab. 1: Plant families with their number of genera and number of species assessed by Next Generation Sequencing (NGS) and optical microscopy.

Figures

Figure 1: A) Rarefaction of genera richness obtained for each honey bee sample with respect to sequencing depth. B) Plot-based comparison of pollen identification through optical microscopy and NGS. Taxonomic assignments are illustrated at the genus level. Positive identification of a taxonomic unit within a sample is indicated in the community matrix as dark gray for microscopy and light gray for NGS. Relative abundance estimations are indicated by size at two levels, i.e. >=5% (fully-filled box) and <5% (half-filled box) of total abundance within a sample. Genera misidentified in optical microscopy were combined for direct comparison and are indicated by quote marks in abbreviated form (Tar = *Taraxacum*, Sco = *Scorzoneroides*, Tan = *Tanacetum*). Availability in the reference database is indicated in the column DB. *For sample 12, three samples were taken from the same study site but different colonies. All three samples were analyzed using NGS to evaluate repeatability, yet optical microscopy was only performed for 12a.

Figure 2: Overall log-scaled relative abundance comparison of genera between the two classification strategies. Rectangles at the axes represent genera only found with one of the two sampling techniques. Pearson's correlation r = 0.86, t = 8.71, df = 26, p < 0.001***.

Figure 3: Pre-clustering evaluation: Starting from the root of the taxonomy of each sequence, every taxonomic level of the assignment was compared to its correct lineage. The overall mean of correct assignments according to the different pre-clustering levels is presented as green dots in the figure (left scale). Similarly, each sequence was tested for erroneous levels of classification with means displayed as red squares and the scale on the right side.

Figure 4: Dependence of sensitivity and specificity by the bootstrap threshold. Sensitivity to identify at species level is illustrated with a red and single-dashed line, whereas generic identification as a red two-dashed line. Specificity is displayed as a green dotted line. The harmonic mean of both species level measures is displayed by a solid black curve, maximized at approximately 0.85 as the suggested optimal classification threshold.